Lecture 7: Estimators and Sampling Distributions

Ignacio Urbina

Introduction & Motivation

- **Objective**: Explore estimators and sampling distributions to assess the probability of obtaining specific estimates, given known distribution assumptions.
- **Why is this important?**
  - To quantify the likelihood of observing a particular estimate or range of estimates.
  - Analyze the behavior of sample estimates relative to population parameters and underlying assumptions.
- This forms the foundation of **statistical inference**, enabling us to draw conclusions about populations from samples.

Population Parameters vs. Sample Estimators

- **Population Parameters**:
  - Fixed values that describe the characteristics of the entire population (e.g., mean μ, variance σ^2).
  - These values are typically unknown and constant (at a given point in time).
- **Sample Estimators**:
  - **Estimator**: A function applied to sample data to estimate population parameters.
  - **Estimate**: A specific value calculated from a sample that serves as an approximation of the population parameter.
  - **Examples**: Sample mean ($\bar{X}$) estimates μ, sample proportion ($\hat{p}$) estimates population proportion, and sample variance (s^2) estimates σ^2.

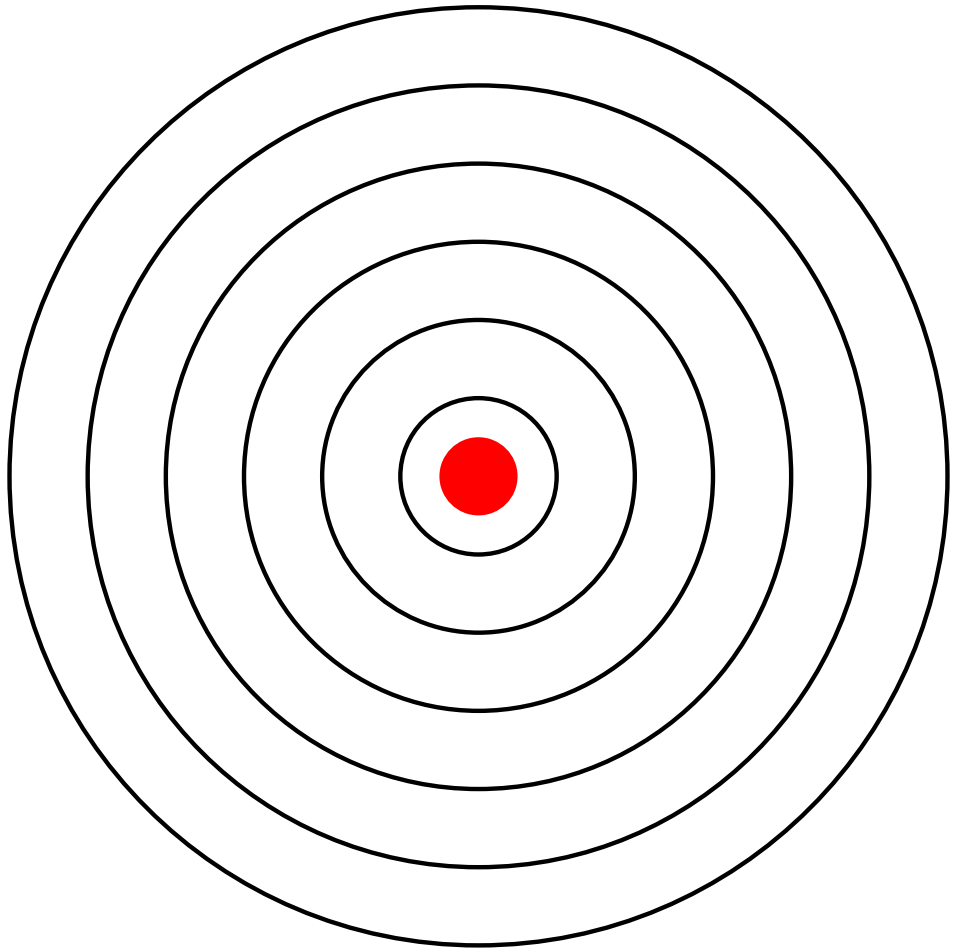
Basic Properties of Estimators

- **Unbiasedness**: The expected value of the estimator equals the population parameter (i.e., $E[\hat{\theta}] = \theta$).
- **Consistency**: As the sample size increases, the estimator approaches the true population parameter.
- **Efficiency**: The estimator has the smallest variance among all unbiased estimators.

Bias and Efficiency

- We define the bias of an estimator $\hat{\theta}$ as:
  - Bias $= E[\hat{\theta}] - \theta$
- Assume two different estimators $\{\hat{\theta}_1, \hat{\theta}_2\}$ of the same population parameter θ.
  - We say that the estimator $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $V(\hat{\theta}_1) < V(\hat{\theta}_2)$
- Ideally, we want to employ an estimator that is unbiased and efficient to approximate a given population parameter of interest.

Bias vs Variance - A Graphical Illustration



Using Estimators to do Inference

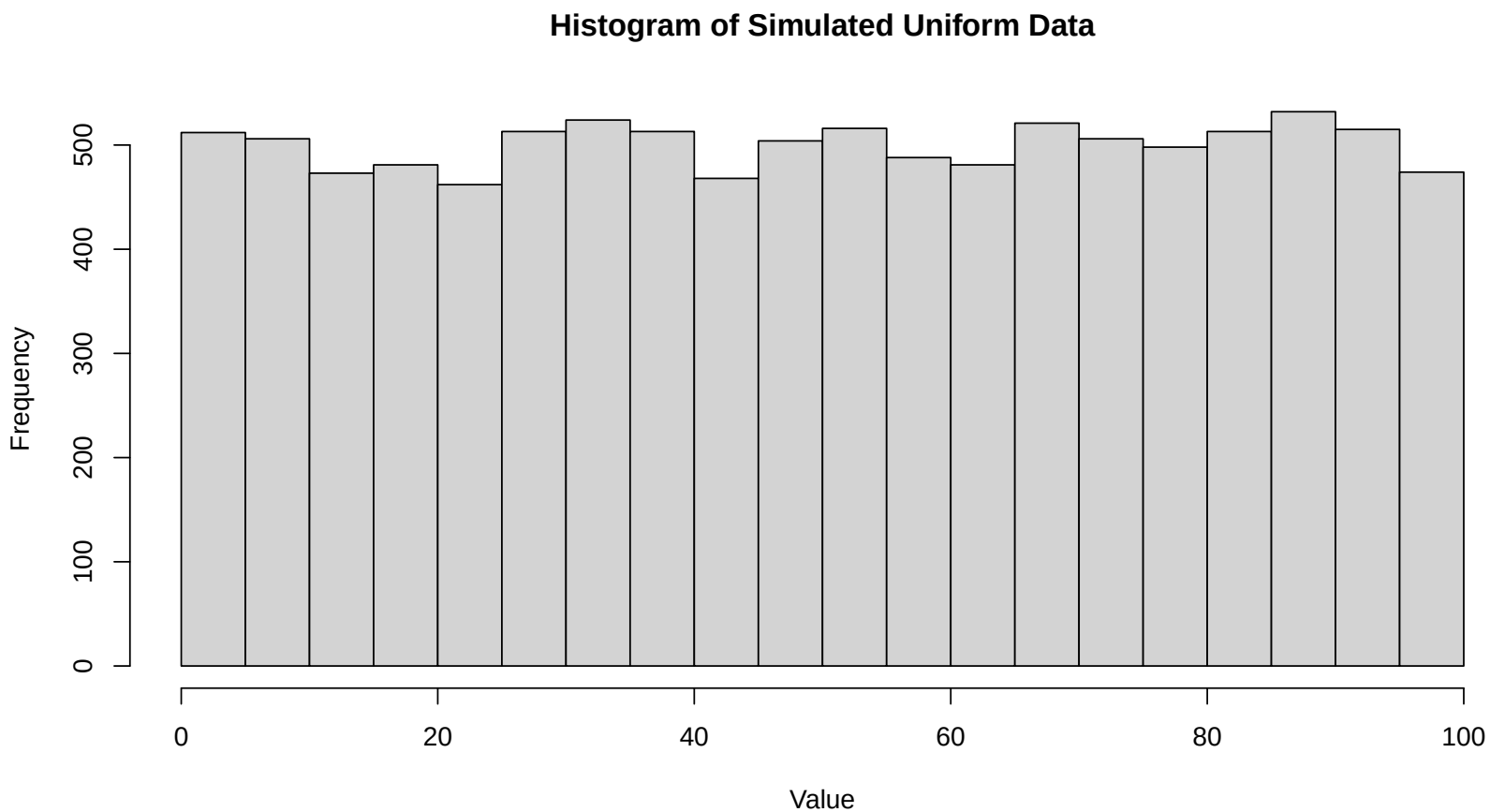
- Say we decide to use an estimator with good properties (i.e., the sample mean). Further, we collect a sample and then compute a specific estimate.
- How can we then **do inferences about the population parameter** (i.e., the true population mean)?
- How can we assess our **degree of confidence** that the specific sample estimate is a good approximation of the true population mean?

Sample Variability - A Simulated Example

- The issue here is **sample variability** or uncertainty. To what extent if we were to collect the sample again we would likely obtain a similar value of the estimate?
- We will do a simulation to illustrate the issue of sample variability.
- **Simulating Uniform(0, 100)**:
  - Generate data from a uniform distribution between 0 and 100.
  - Population size: 10,000.

```
set.seed(789)  # Set seed for reproducibility
pop_data <- runif(10000, min = 0, max = 100)  # Generate 10,000
    random values from a uniform distribution
# Plot histogram of the uniform data
hist(pop_data, breaks = 30,
  main = "Histogram of Simulated Uniform Data",
  xlab = "Value")
```

Plotting the *Population* Data



Random Sampling from Simulated Uniform Data

- Wel'll draw **three random samples from the simulated population (Uniform(0, 100))**:
  - Sample sizes: $N = 20$ (three different samples).
  - Compute the sample mean for each case.

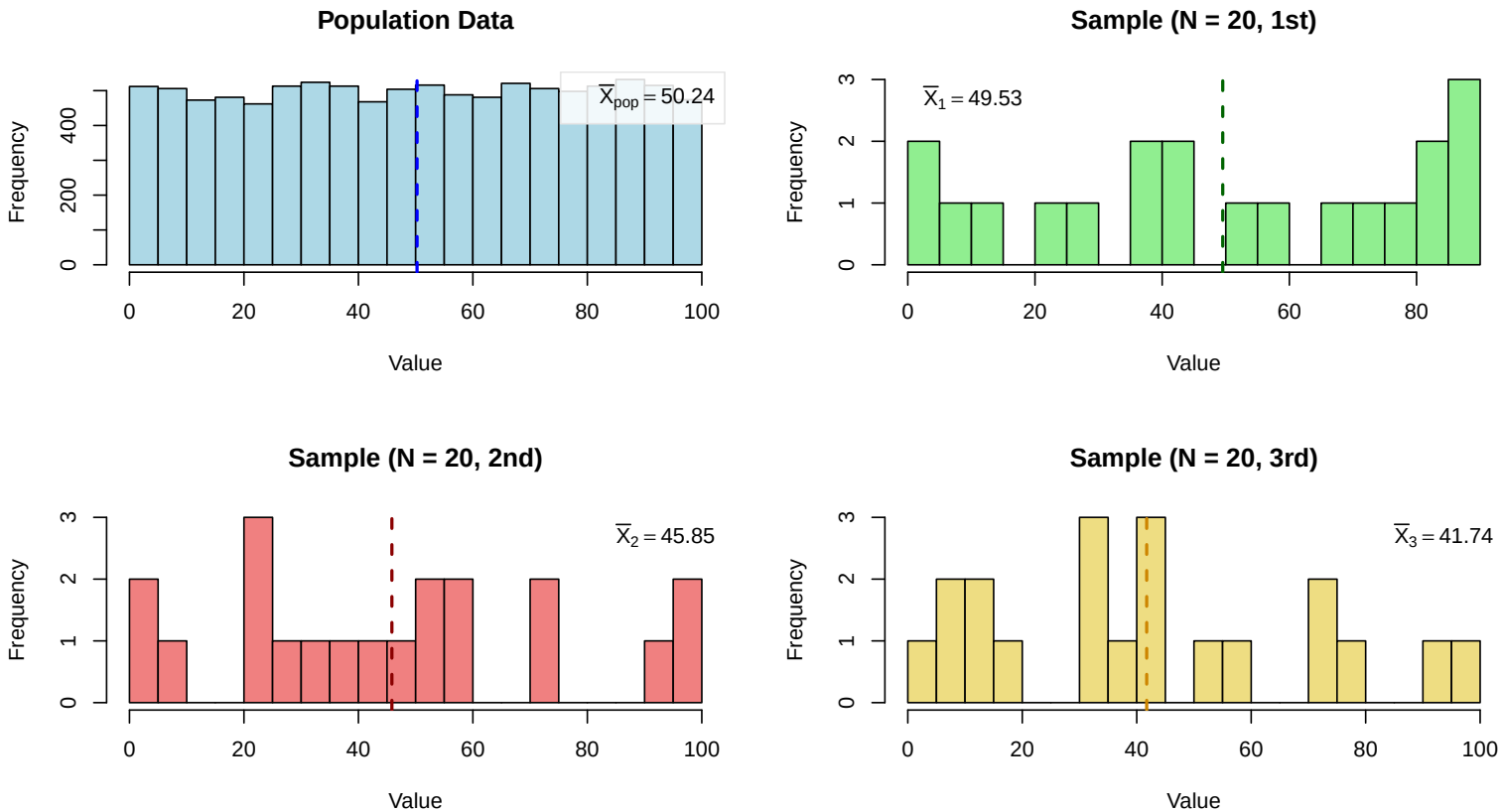
```
# Population mean
true_mean <- mean(pop_data)  # Calculate the true mean of the
    population

# Random samples from the population
set.seed(4321)  # Set seed for reproducibility
sample_20_1 <- sample(pop_data, size = 20)  # First random sample of
     size 20
sample_20_2 <- sample(pop_data, size = 20)  # Second random sample
    of size 20
sample_20_3 <- sample(pop_data, size = 20)  # Third random sample of
     size 20

# Compute sample means
mean_20_1 <- mean(sample_20_1)  # Mean of the first sample
mean_20_2 <- mean(sample_20_2)  # Mean of the second sample
mean_20_3 <- mean(sample_20_3)  # Mean of the third sample
```

Computing the True Mean and Sample Means

- **True mean of the population**:
  - 50.24
- **Sample means**:
  - $N_1 = 20$, $\bar{X}_1 = 49.53$
  - $N_1 = 20$, $\bar{X}_2 = 45.85$
  - $N_1 = 20$, $\bar{X}_3 = 41.74$

Visualizing the Sampling Process

- We can observe differences in the shape of the distribution across the samples and compared to the population distribution.



Understanding Sampling Variability

- **Imagine we had only done the process once** and obtained a sample mean equal to 49.53. How much confidence should we place on that estimate?
- Our small exercise shows some **variability**: taking the sample again changes the estimate.
- So, given the sample size, how likely is it that we would get a wide range of different values or a small one?
- To answer this, we will introduce the concept of **sampling distributions**.

Estimators and Sampling Distributions

- **Sampling Distribution**:
  - The probability distribution of an estimator.
  - **Key Concept**: Sampling distributions allow us to understand the variability of estimators.
- **Example**: Distribution of the sample mean $\bar{X}$ when repeatedly sampling from the population.

Example: Sampling Distribution for a Normal Population

- Assume a continuous random variable $X \sim N(\mu, \sigma^2)$ – aka "*a normal population*." Assume we draw a sample of size n.
- What is the sampling distribution of the sample mean $\bar{X}$?
- We can start answering this question by deriving $\bar{X}$ true mean and standard deviation (also known as standard error).
- Following the derivation included in last week's lecture:
  - $E[\bar{X}] = \mu$
  - Standard Error: $SE(\bar{X}) = SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Standard Error vs Standard Deviation

- **Standard Deviation (σ)**:
  - Measures the spread of a population.
- **Standard Error (SE)**:

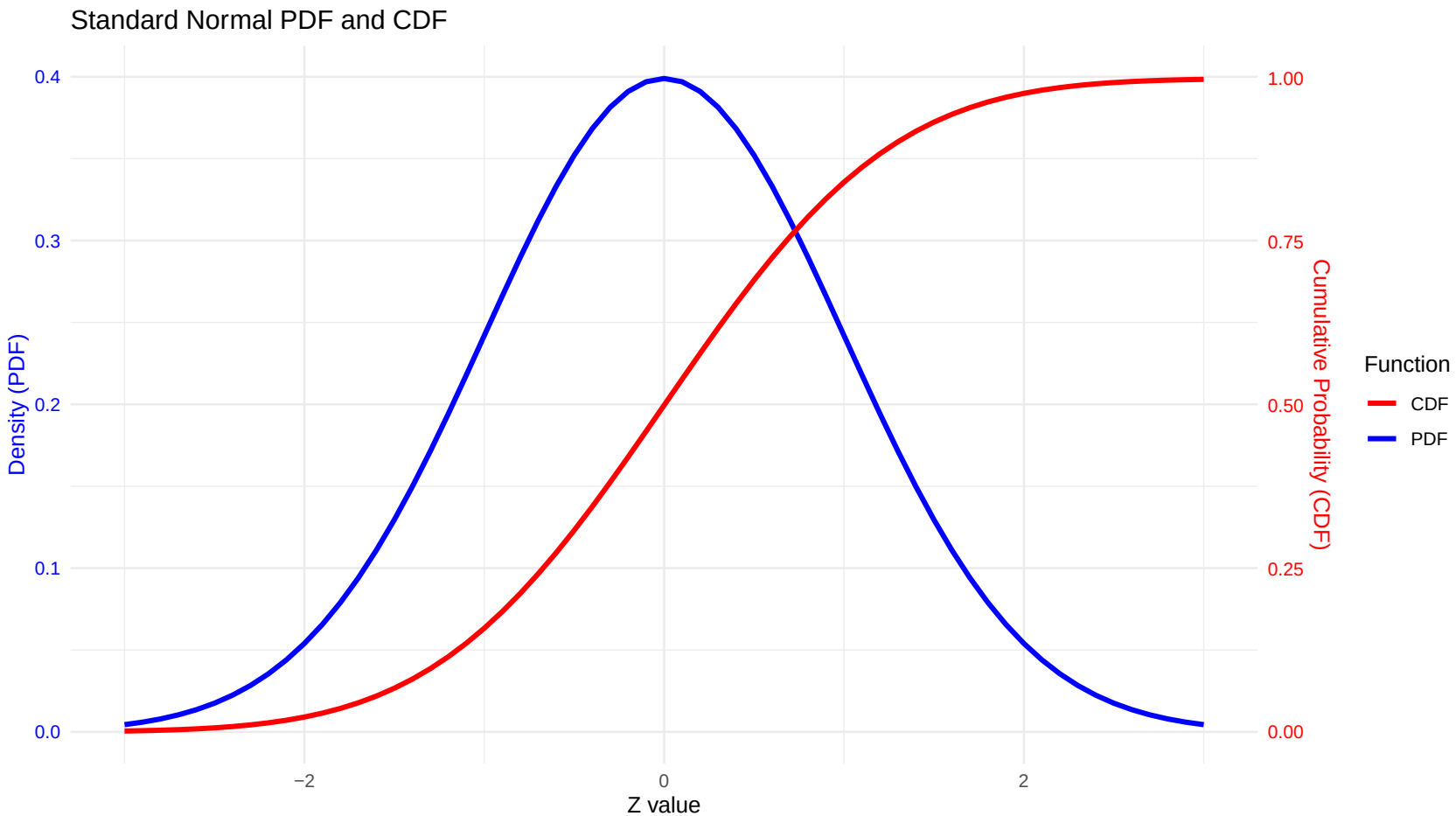
$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

  - Measures the spread of an estimator (e.g., the sample mean).
  - Simply put, the Standard Error is the standard deviation of an estimator (recall that estimators *are random variables*)
- **Difference**: SE decreases as the sample size increases, reflecting more precise estimates.

Detour: Quick Review of the Normal Distribution

- **Probability Density Function (PDF)**:
  - $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
  - Where μ is the mean and σ is the standard deviation.
- **Properties**:
  - The mean, median, and mode are all equal.
  - Symmetrical PDF.
  - Approximately 68% of the data falls within 1 standard deviation of the mean, 95% within 2, and 99.7% within 3 (aka, **the empirical rule**).
- **Standard Normal Distribution**: A normal distribution with mean 0 and standard deviation 1, often denoted by $Z \sim N(0, 1)$.

Plotting PDF and CDF of Standard Normal Distribution



Cumulative Distribution Function (CDF) of Standard Normal

z	CDF	z	CDF	z	CDF	z	CDF
-3.00	0.00	-1.40	0.08	0.20	0.58	1.80	0.96
-2.80	0.00	-1.20	0.12	0.40	0.66	2.00	0.98
-2.60	0.00	-1.00	0.16	0.60	0.73	2.20	0.99
-2.40	0.01	-0.80	0.21	0.80	0.79	2.40	0.99
-2.20	0.01	-0.60	0.27	1.00	0.84	2.60	0.99
-2.00	0.02	-0.40	0.34	1.20	0.88	2.80	1.00
-1.80	0.04	-0.20	0.42	1.40	0.92	3.00	1.00
-1.60	0.06	0.00	0.50	1.60	0.94		

Table 1: CDF of Standard Normal Distribution, $CDF(z) = \Pr(Z < z)$

Computing Probabilities Using the Standard Normal

- **Computing the Probability** $P(Z < 1) = \mathbf{0.84}$:

```
P_Z_less_than_a <- pnorm(1)  # pnorm(a) is a R function that
    computes CDF(a) for the std. normal distr.
P_Z_less_than_a
```

```
## [1] 0.8413447
```

- **Computing the Probability** $P(Z < -0.8) = \mathbf{0.21}$:

```
P_Z_less_than_b <- pnorm(-0.8)  # Compute the prob. that Z<-0.8
P_Z_less_than_b
```

```
## [1] 0.2118554
```

- **Probability** $P(-0.8 < Z < 1) = \mathbf{0.63}$:

```
P_between_a_and_b <- pnorm(1) - pnorm(-0.8)  # Compute the prob.
    that Z is between -0.8 and 1
P_between_a_and_b
```

```
## [1] 0.6294893
```

Z-Scores & Probability Computations

- **Z-Score Formula**:
$$Z = \frac{X - \mu}{\sigma}$$

- **Theorem 1**: If $X \sim N(\mu, \sigma^2)$, then Z follows a standard normal distribution $N(0, 1)$.
  - **Takeaway**: If we can **assume** $X \sim N(\mu, \sigma^2)$, then we can standardize any value of X using the Z-score to compute probabilities under the standard normal distribution.

- **Example**: Given $X \sim N(50, 25)$, compute $P(X < 55)$ using the standard normal CDF table:
  - Note $\sigma = \sqrt{\sigma^2} = \sqrt{25} = 5$.
  - Convert $X = 55$ to $Z \longrightarrow Z = \frac{X-50}{5} = \frac{55-50}{5} = 1$
  - Use a Z-table to find the corresponding probability, $P(X < 55) = P(Z < 1) = 0.84$.

Z-Scores & Probability Computations

- Given $X \sim N(\mu, \sigma^2)$, how do we use the standard normal CDF table if we want to compute $P(X > b)$?
- **Example 2**: Given $X \sim N(50, 25)$, compute $P(X > 42)$ using the standard normal CDF table:
  - Note $\sigma = \sqrt{25} = 5$.
  - Convert $X = 42$ to Z: $Z = \frac{42-50}{5} = -1.6$
  - Use a Z-table to find the probability of the complement of $X > 42$: $P(X < 42) = P(Z < -1.6) = 0.06$,
  - Then use the complement rule: $P(X > 42) = 1 - P(X < 42) = 1 - P(Z < -1.6) = 1 - 0.06 = 0.94$

Wrapping Up Normal CDFs $\longrightarrow$ Back to Sampling Distributions

- We took a short **detour** to review the **standard normal distribution** and how to:
  - Compute probabilities using the **CDF of** $Z \sim N(0, 1)$
  - Convert values from $X \sim N(\mu, \sigma^2)$ into **Z-scores**
  - These tools are essential for doing inference with **normally distributed estimators**.
- Now, let's return to our core question:
  - What is the distribution of the **sample mean** $\bar{X}$ when each $X_i \sim N(\mu, \sigma^2)$?
- Key observation:
  - $\bar{X}$ is a **linear combination (or scaled sum)** of the individual X_i's.
  - So, to fully understand $\bar{X}$, we first need to understand: The distribution of **sums of normal random variables**.

Distribution of a Sum of Normal RVs

- Assume there is a sequence of independent draws of Y, denoted by y_i, all following a normal distribution $Y \sim N(\mu, \sigma^2)$. Define the sum of these draws as:

$$S = \sum_{\forall i} y_i = y_1 + y_2 + \cdots + y_N$$

- What is the distribution of S?
- **Theorem 2**: If $Y_1, Y_2, \ldots, Y_N$ are independent and identically distributed random variables with $Y_i \sim N(\mu, \sigma^2)$, then the sum S also follows a normal distribution:

$$S \sim N(E[S], V[S])$$

  - This applies to any linear combination of the form: $S = \sum_{\forall i} a_i \cdot y_i$ in which a_i corresponds to fixed coefficient (not RVs).

Sampling Distribution of $\bar{X}$ when $x_i \sim N(\mu, \sigma^2)$

- Assume there is a series of independent draws of X, denoted by x_i, all following a normal distribution $X \sim N(\mu, \sigma^2)$.
- Define the sample mean as a linear combination:

$$\bar{X} = \frac{\sum_{\forall i} x_i}{N} = \frac{x_1}{N} + \frac{x_2}{N} + \cdots + \frac{x_N}{N}$$

- What is the distribution of $\bar{X}$?
- **Theorem 3**: Applying the previous theorem: $\bar{X} \sim N(E[\bar{X}], V[\bar{X}])$. Thus,

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{N}})$$

Sampling Distribution of the Sample Mean for a Normal Population

- Assume $X \sim N(\mu = 50, \sigma^2 = 25)$. Then, **compute the probability** $P(\bar{X} > 51)$ assuming a sample size of $n = 100$ using a standard normal CDF table:

```r
# Given mean (mu), standard deviation (sigma), and sample size (n)
mu <- 50  # Population mean
sigma <- 5  # Population standard deviation
n <- 100  # Sample size
b <- 51  # Threshold value for sample mean
# Standard error
SE <- sigma / sqrt(n)  # Calculate the standard error of the sample
    mean
# Z-score
Z_b <- (b - mu) / SE  # Calculate the Z-score for the sample mean
    threshold
# Probability P(bar(X) > b)
P_X_greater_than_b <- 1 - pnorm(Z_b)  # Calculate the prob. that the
     sample mean > b
P_X_greater_than_b
```

```
## [1] 0.02275013
```

Recap, Sampling Distributions

- Starting with a known distribution for a random variable under examination, we can derive the exact sampling distribution of the sample mean.
- The sampling distribution can be used as a tool to measure the underlying uncertainty of a given estimate, which allows us to **perform statistical inference**.
- Yet, in most cases, we do not know the underlying distribution of a random variable we are studying (i.e., the population distribution is unknown).
- *How can we then perform statistical inference with a sampling estimate when the population distribution is unknown*?

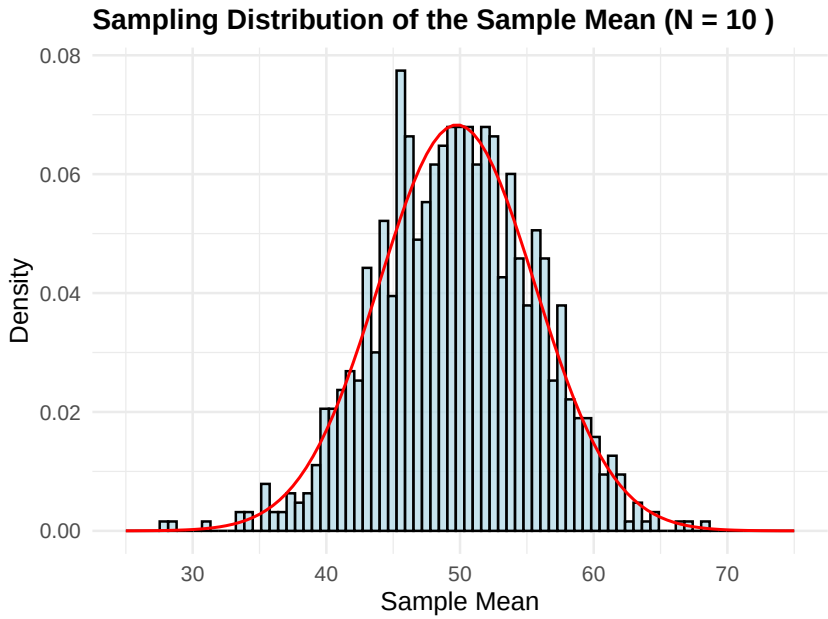
The Central Limit Theorem (CLT)

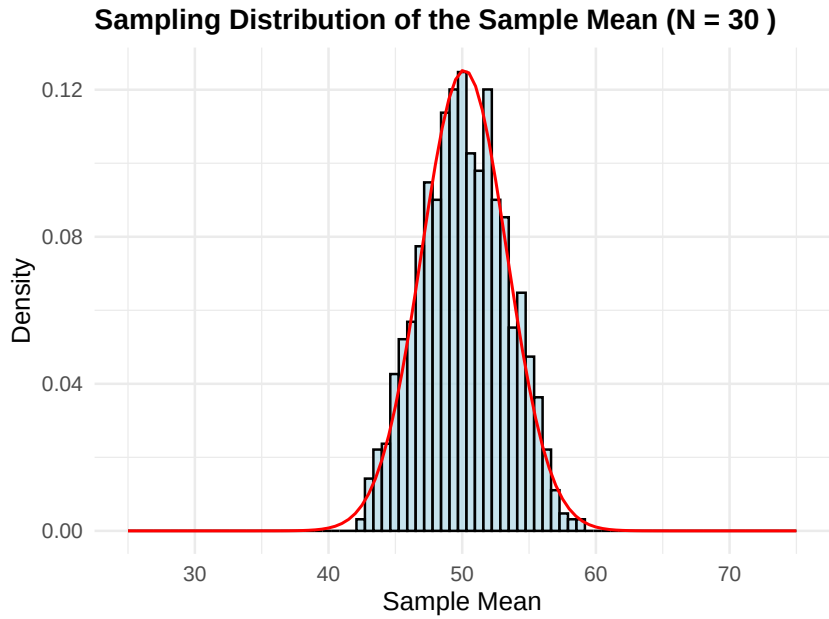
- **Theorem**: For a sufficiently large sample size, the sampling distribution of the sample mean $\bar{X}$ is approximately normal, *regardless* of the population's distribution.
- **Key Implication**: This enables us to use *Z-score transformations* to approximate probabilities for $\bar{X}$, even when the original variable X is not normally distributed.
- **Why It Matters**: The CLT justifies inference on population parameters using sample data, making it foundational for hypothesis testing and confidence intervals.

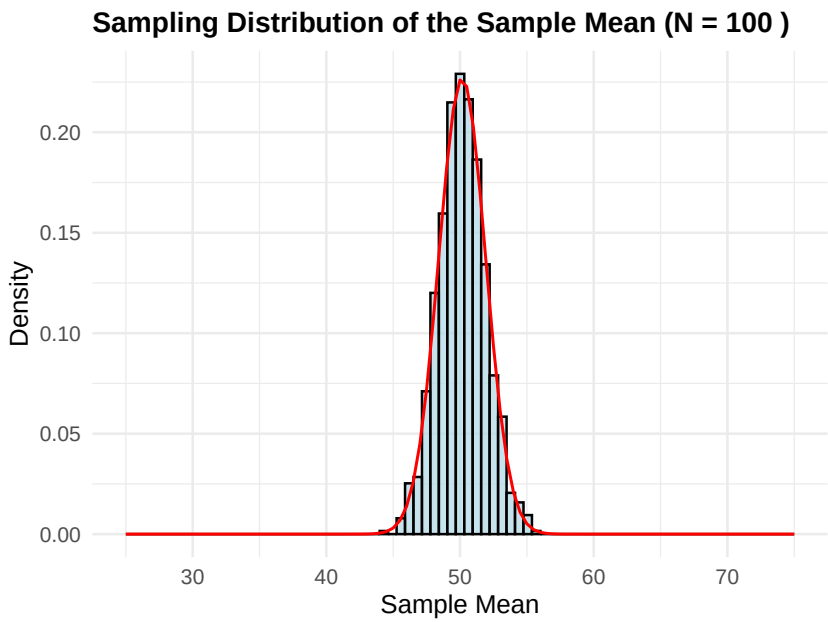
Assumptions of the Central Limit Theorem (CLT)

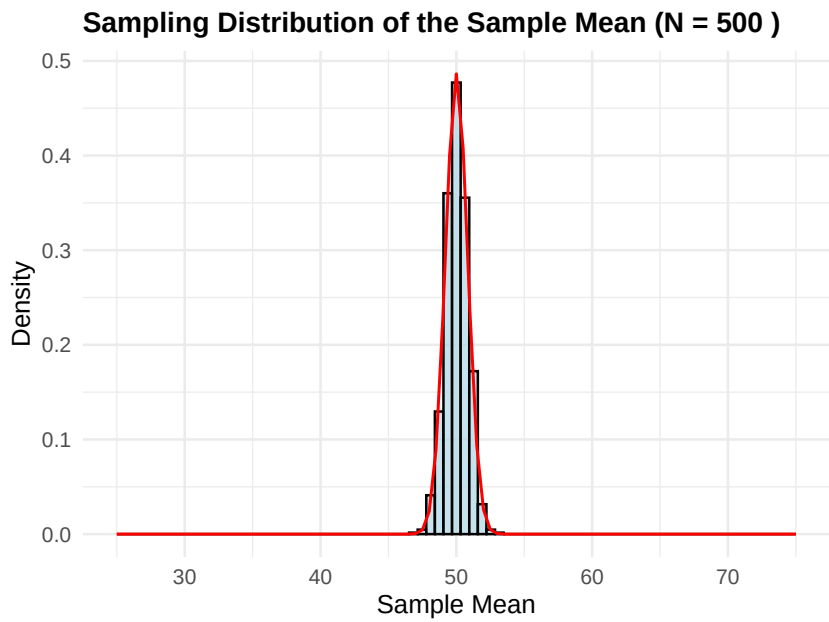
- For the CLT to hold, the following conditions must be met:
  - The sample consists of **independent** observations
    - Simple Random Sampling (SRS) supports the independence assumption
  - The observations are drawn from the **same distribution**
  - The underlying distribution has a **finite mean** and **finite variance**
  - The **sample size is sufficiently large**
- Notes:
  - "Sufficiently large" depends on the **shape of the population**:
    - For populations with symmetric distributions: $n \geq 30$ often works well
    - Skewed or heavy-tailed distributions may require larger n

Simulated Example of CLT (Population is *Uniform*(20, 80))

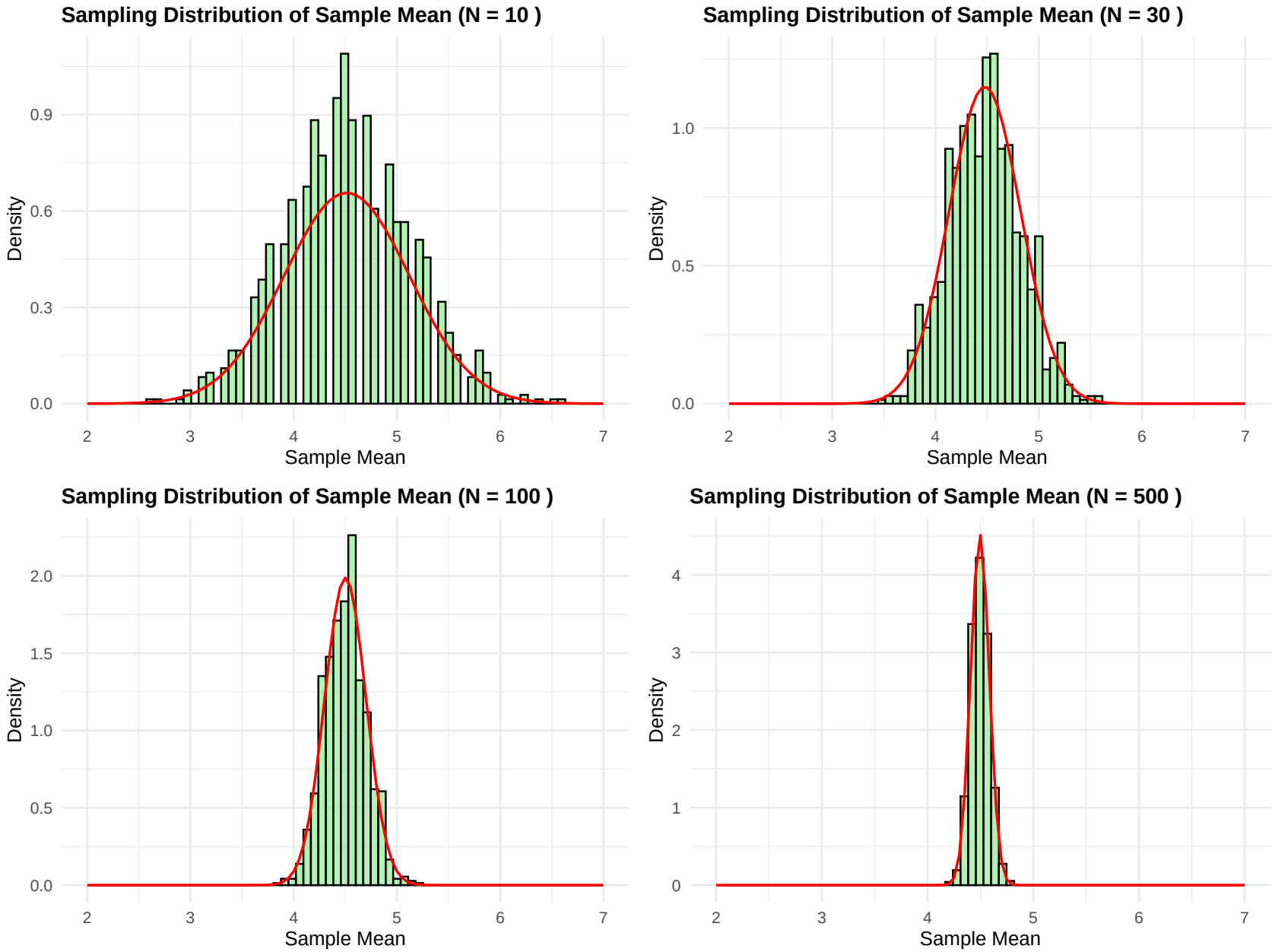








CLT Example 2 (Pop. is $Binomial(n = 20, p = 0.15)$)



Z-Scores, the CLT, & Probability Computations

- Invoking the CLT **assume** $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$:
  - Use the Z-score formula to standardize the sample mean $\bar{X}$ for probability computations.
  - **Z-Score Formula**:
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$
- **Example**:
  - Given $\bar{X} \sim N(50, \frac{25}{100})$ (where $\mu = 50, \sigma^2 = 25, n = 100$), compute $P(\bar{X} > 51)$.
  - Convert to Z:
$$Z = \frac{51 - 50}{\frac{5}{\sqrt{100}}} = 2$$
  - Use a Z-table to find the probability corresponding to $Z = 2$.

From Theory to Practice: The Plug-in Principle & the CLT

- For a large n, the Central Limit Theorem assumes:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  where μ and σ are **population parameters** (usually unknown).

- In practice, we don't know these parameters.
  - We apply the **plug-in principle**: use sample statistics to estimate unknown parameters. This is justfied by the Law of Large Numbers (LLN).

- Replace:
  - μ with $\bar{Y}$ (sample mean)
  - σ with s (sample standard deviation)

- This gives the approximation:

$$\bar{Y} \approx N\left(\bar{Y}, \frac{s^2}{n}\right)$$

The Law of Large Numbers (LLN)

- **Law of Large Numbers** (LLN): As the sample size n increases, the *sample mean* $\bar{X}_n$ converges to the *population mean* μ.
  - More formally: If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with finite mean μ, then:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu \quad \text{as } n \to \infty$$

  (Convergence in probability)
- **Key assumptions**. The observations must be:
  - *Independent*
  - *Identically distributed (i.i.d.)*
  - With a *finite expected value* $\mathbb{E}[X] = \mu$
- **Why it matters**: With enough independent data, the sample mean gets arbitrarily close to the true mean.

Using the CLT - Practical Example.

Suppose we have a random sample of a random variable y_i of sample size $N = 400$. Further, suppose $\bar{Y} = 23.5$ and $s = 20$ (sample sd). What is the probability that the sample mean is higher than 25?

```
# Given values
mean_y <- 23.5  # Sample mean
s <- 20  # Sample standard deviation
N <- 400  # Sample size
# Calculate the standard error of the mean
se_y <- s / sqrt(N)
# Calculate the probability directly
prob_calc <- 1 - pnorm(25, mean = mean_y, sd = se_y, lower.tail =
    TRUE)  # lower.tail = TRUE gives P(X < x)
# Display the probability
prob_calc
```

```
## [1] 0.0668072
```

- **Result**: *If we draw a new sample of same size*, the probabiltiy that the sample mean is greater than 25 is app. 6.7%.

Conclusion

- **Key Takeaways**:
  - Estimators help us understand population parameters using sample data.
  - Sampling distributions allows to measure uncertainty in a given estimate.
  - The **Central Limit Theorem** is foundational for statistical inference.
  - Z-scores are useful for probability calculations under normal assumptions or when we invoke the CLT.